Homogeneity Assessment of Phenological Records from the Swiss Phenology Network

Yuri Brugnara[1], Renate Auchmann[1], This Rutishauser[1], Regula Gehrig[2], Barbara Pietragalla[2], Michael Begert[2], Christian Sigg[2], Valentin Knechtl[2], Thomas Konzelmann[2], Bertrand Calpini[2], Stefan Brönnimann[1]

[1]Oeschger Centre for Climate Change Research and Institute of Geography, University of Bern, Bern, Switzerland

[2] Federal Office of Meteorology and Climatology, MeteoSwiss, Zurich, Switzerland

Corresponding author

Yuri Brugnara

Email: yuri.brugnara@giub.unibe.ch

Tel.: +41 (0)31 631 85 24

ORCID: 0000-0001-8427-0064

Acknowledgements

This work was funded by MeteoSwiss in the framework of GCOS Switzerland.

source: https://doi.org/10.7892/boris.133017 | downloaded: 5.5.2023

Abstract. Phenological data have become increasingly important as indicators of long term climate change. Consequently, long-term homogeneity of the records is an important aspect. In this paper we apply a breakpoint detection algorithm to the phenological series from the Swiss Phenology Network (SPN). A combination of three statistical tests is applied and different constraints are tested with respect to the choice of reference series. Breakpoint detection is only possible for a fraction of the series due to the shortness of some series and the lack of suitable reference series. Spring phases are more likely to be suitable than fall phases because of their higher spatial correlation. Out of nearly 3,000 phenological series with at least 20 data points, only about 5% were found to be significantly inhomogeneous, although a visual validation indicates that many mid-sized breakpoints remained undetected. The detected breakpoints were compared with metadata and more than half of them could be attributed to a change of observer.

Keywords: plant phenology, breakpoints, Switzerland, climate change

1. Introduction

Observations of plant phenological phases not only constitute a monitoring of plant life, but serve the assessment of agricultural suitability, changes in habitat factors and others. Among the factors affecting changes in plant phenology, climate is one of the most important. This is particularly the case in spring. Because of high temperature sensitivity of many phenological phases, plant phenology has become an important climate change impact indicator in Switzerland (Studer et al. 2005; Seiz and Foppa 2007), in Europe (Menzel et al. 2006; Fu et al. 2015) and globally (Cramer et al. 2014), although warming effects are potentially not stable over time (Rutishauser et al. 2008; Fu et al. 2015). Its independence from instrumental temperature measurements makes phenology a particularly attractive indicator of global warming (Anderson et al. 2013). Furthermore, because observations of plant phenological phases date back several centuries, phenological observations can be used as a proxy for climate reconstruction (Rutishauser et al. 2008; Ge et al. 2014).

For these reasons, phenology has been defined a relevant parameter for the National Climate Observing System (GCOS Switzerland) (Seiz and Foppa 2007; MeteoSwiss 2018), and is as such recognized an important factor in climate monitoring for terrestrial observations of the biosphere. However, using phenological observations for assessing climate trends brings specific requirements with respect to long-term stability (Dose and Menzel 2004; Seiz and Foppa 2007; Schleip et al 2008; WMO 2016). The stability of the observing configuration in phenology is dependent on how long the same observer is active, on the change of the observed plants, on changes in the environment or in the plant itself and on the quality of the observation instructions. Undesired signals in the time series must be flagged by detecting so-called breakpoints, i.e. points in time where a significant change occurred in the observing procedure (Dose and Menzel 2004; Rutishauser et al 2009; Keatley and Hudson 2012).

While quality control procedures for phenological data have been developed (Hense and Müller 2007) and are routinely applied, breakpoint detection methods are only rarely used in plant phenology (e.g., Dose and Menzel 2004; Schleip et al 2008). Conversely, while breakpoint detection methods are normally used in satellite phenology (e.g., Verbesselt et al 2010; Jamali et al 2015), series of plant phenological observations share more similarities with meteorological series. In this paper we apply breakpoint detection methods typically used in climatology to phenological series from the Swiss Phenology Network (SPN: Defila and Clot 2001).

Concerning meteorological data like temperature and precipitation, breakpoint detection methods are routinely applied (e.g., Begert et al. 2005; Kuglitsch et al. 2012). In a recent COST Action, detection and homogenisation methods were compared, and benchmark data sets have been constructed that allow rigorous testing (Venema et al. 2012; see also Willett et al. 2014). With respect to breakpoint detection, we recently have applied a combination of tests to the meteorological series from Switzerland (Kuglitsch et al. 2012). Here we analyse to what extent these methods are also suitable for phenological series from the same region.

The paper is organised as follows. In Section 2 we describe the data, the breakpoint detection methods, and the evaluation strategy of the detection approaches. In Section 3 we show the results of applying the methods. In Section 4 we discuss the limitations of applicability of the approach and compare the results to those typically obtained for meteorological series. Conclusions are drawn in Section 5.

2. Materials and Methods

2.1. The Swiss Phenology Network

This paper uses the data from the Swiss Phenology Network (SPN) which was initiated in 1951 (Defila et al. 2016) with 70 stations. The SPN comprises today 167 stations across Switzerland (Fig. 1). The onset dates of up to 69 different phenological events (for 26 different species) are currently being observed. Twenty-eight of these phenological events have been observed since the beginning, whereas the other 41 started in 1996 within a renewal of the observation program (Defila 2008). A dataset from 1951 to 2015 was used, comprising 9,455 series with about 200,000 single observations. However, in this study only 2,925 series were tested, although 7,393 series had been used within experiments to find the best test configuration (see Table 1).

The data underwent careful quality assurance, which is described in detail in Auchmann et al. (2018). In this process, suspect data were flagged according to a sequence of checks that was based on range, internal consistency, biological plausibility, and comparisons with the same series from neighbouring stations or different series from the same station. For this study, the flagged data were not used.

The metadata we used in this study are the years of the observer changes. The observers of the Swiss Phenology Network are volunteers who observe the plants in their neighbourhood following instructions from MeteoSwiss and a detailed manual (Brügger and Vassella 2003). A change of an observer often implies a change of the observed plants, which can cause breaks in the data series of that station.

2.2. Breakpoint detection methods

Historical phenological records, just like meteorological ones, are prone to inhomogeneities caused by factors that affect the way observations are carried out (e.g., a change of the observer). Inhomogeneities can be detected using statistical methods, especially if they happened at a specific time point (so-called “breakpoints”). Knowing which data series are affected by significant breakpoints is particularly important when analysing trends.

We used an algorithm similar to that used for Swiss temperature series in Kuglitsch et al. (2012). We independently applied three statistical tests to each phenological series that has at least 20 observations (shorter series constitute a too small sample for meaningful statistical testing). The agreement among the three tests determines which breakpoints are to be considered significant. The main advantage of this approach is the reduction of false detections. The tests, however, are not the same as used in Kuglitsch et al. (2012).

Each test is applied to difference series between the candidate and well-correlated reference series. The whole procedure is fully automatic and reproducible, the detection does not involve subjective decisions after the initial parameters are set. The tests that we used are the following:

− standard normal homogeneity test (SNHT; Alexandersson and Moberg 1997);

− Pettitt's test (Pettitt 1979);

− penalized maximal t test (PMT; Wang 2008).

The SNHT and Pettitt's test are based on rather simple statistical tests in which a series is divided into two parts. The null-hypothesis is that the means of the two sub-samples are identical. The main difference between SNHT and Pettitt's is that the latter is non-parametric and as such it does not require a normal distribution. These tests can detect only a single breakpoint at a time, therefore they are applied recursively to homogeneous sub-periods as long as these have at least 20 data points.

The PMT test is an improvement of the SNHT test in which a penalty factor is introduced to reduce the false alarm rate at the edges of a series (i.e., near the beginning and the end). This implies a better hit rate in the middle but a lower hit rate at the edges (see also Wang et al. 2007). The PMT test was applied using an adapted version of the R software RHtestsV5.

The p-value threshold for significance was set to 0.05 in all three tests. A breakpoint is considered detected by a certain test if the test finds it in at least three difference series (candidate minus three reference series). We allowed a tolerance of one year (for example, if the first difference series has a breakpoint in 1979, the second in 1980, and the third in 1981, then all of them have a breakpoint in 1980). If two or three tests detect the same break (always +/- 1 year) a breakpoint is set ("significant" breakpoint). In general, breakpoints are merged with an iterative procedure that starts with 0 tolerance and increases to maximum tolerance (one year). This way the year with more detections is preferred.

The statistical tests that we used have been developed mainly for temperature and precipitation, although they have been used in other areas (e.g., Čufar et al. 2012). Phenological and temperature series have similar statistical properties. For instance, annual temperature means are normally distributed; 86% of the Swiss phenological series with at least 20 observations do not differ significantly ($p \leq 0.05$) from a normal distribution, according to the Lilliefors (Kolmogorov-Smirnov) test. However, similarly to precipitation, phenological series are in general less correlated in space than temperature (Güsewell et al. 2018), which makes the detection of breakpoints less effective. This is mainly because the behaviour of plants is more complex than that of temperature, being dependent on both meteorological and biological factors. Temperature series (annual means) usually have correlations larger than 0.9 when distances between stations are in the order of tens of kilometres, whereas many phenological series do not reach correlations above 0.7 for the same distances.

Phenological series, like precipitation, also show larger inter-annual variability in comparison with temperature, which negatively affects the signal-to-noise ratio of inhomogeneities. For these reasons, we expect the breakpoint detection in phenological series to have a hit rate similar to those estimated for precipitation, i.e. in the range of 5-25% (Venema et al. 2012).

2.3. Reference series

Kuglitsch et al. (2012) used ten reference series for temperature (of which at least 3 must detect the same breakpoint for it to be considered significant). For phenological series, a potential issue of false detections arises because single plants can react rather abruptly to environmental factors. To reduce false detections, we used a smaller number of reference series. This has the side effect of reducing the hit rate of the algorithm, particularly for midsize breakpoints, but also has the advantage of increasing the quantity of data for which it is possible to find enough reference series to perform the breakpoint detection.

Requiring the same number of reference series for each and every target series guarantees some consistency in the method (i.e., the probability of finding a breakpoint is similar everywhere). If this is possible only for a sub-period of a series (i.e., by picking a later starting year), then the breakpoint detection is performed only on that sub-period (this affects 29% of the analysed series).

We tested five different approaches of how to select reference series. The approaches made use of information such as phase/species, altitude, correlation, overlap, and tolerance year. The experiments differ in the combination of selection criteria (except minimum correlation, which was always set to 0.6). Table 1 shows an overview of the experiments.

Experiments ALL1 and ALL2 did not use biological constraints. We used eight reference series with one (ALL1) or two (ALL2) years tolerance for the detected breakpoints among the three different tests. The use of any reference series that is well correlated with the candidate, independently from its plant species or phase, had the advantage that enough reference series could be found for every series in every period. The disadvantage was that some data series ended up with reference series with spurious correlation that are unsuitable for comparisons in a biological context, for example spring phases were correlated with autumn phases, which react differently to climate factors.

Therefore, in experiment 8REF we introduced the biological constraint (and a maximum difference in onset days of 30 days, again 8 references were used). Inter-species correlations are often as large as those between the same species. Using other species as reference, however, would increase the risk of misinterpreting different biological reactions to forcings such as a rapid warming. Additionally, in 8REF a reference series could not come from the same station of the candidate series unless no other alternatives were available, to avoid simultaneous inhomogeneities due to changes of observer. The number of tested series dropped in this experiment due to the lower amount of potential reference series.

To enlarge the number of series that could be tested and at the same time reach a low false detection rate, we tested experiment 5REF using five reference series (experiment 5REF_NOQC, which is the same as 5REF but using not quality controlled data, showed fewer breakpoints due to larger noise and was not considered further). In the 5REF experiment, a breakpoint has to be seen by the majority of the references (three out of five) to be confirmed. Therefore, it represents the more conservative setting. In the remainder of the paper, unless noted otherwise, results for 5REF are shown.

2.4. Use of metadata

The years when changes of observer occurred are used to adjust the position of detected breakpoints (metadata adjustment): if a breakpoint is detected one year before or after the year when the observer changed, it was

moved to the year of the change. Metadata adjustment was done for each of the three tests separately and again on the final set of breakpoints. In a similar fashion, breakpoints were forced to be at the year preceding a large gap (≥3 years) if a statistically detected breakpoint appears one or two years before the first gap year or if it appears in the first year of observation after the gap.

The highest fraction of breakpoints that are confirmed by metadata is found for the 5REF experiment (Table 1). This is an indication that 5REF has, as expected, the lowest false alarm rate.

2.5 Verification

A proper validation of the breakpoint detection algorithm would require a benchmark consisting of surrogate series. Such benchmarks have only recently been developed in the climate sciences (e.g., Venema et al. 2012) and require detailed knowledge of the causes of the breakpoints and of their statistical properties. Two of the three tests used in this paper have been benchmarked in Venema et al. (2012); however, breakpoints in phenological series are arguably rather different from those affecting temperature series. For instance, events such as parasites attacking a plant do not have a correspondence in meteorological data.

In the absence of surrogate series, we only provide a subjective validation, based on the visual analysis of a randomly selected sub-sample of series, using standardised summary plots produced by the breakpoint detection software (see Fig. 5). We inspected all the series found inhomogeneous, plus 100 randomly selected homogeneous series. The inspection was carried out independently by three of the authors (hereafter referred as "experts").

3. Results

3.1. Correlation with reference series

Figure 2 shows the distribution of the correlations of the reference series as a function of the phenological phase. The highest correlations are found for the start of flowering and for vintage, the lowest for hay harvest, leaf colouring and leaf drop. Figure 3 shows the correlations as a function of elevation difference. As shown by the red line, more than half of all reference series used in the whole data set were drawn from stations no more than 125 meters higher or lower than the candidate station. Correlations are higher when the elevation difference is below 125 m (median of 0.7). However, once the elevation difference is larger than 250 m, elevation difference becomes no longer important. The figure also shows that the spread of correlation for different phenological phases becomes smaller once the elevation difference becomes larger. This is also because the reference series are constrained to a minimum correlation of 0.6.

There are no appreciable differences among species in the correlation changes with the elevation (not shown).

3.2. Feasibility of the breakpoint detection

Not for all series we find sufficient (five) references series that fulfil the conditions specified in Table 1. For the case of 5REF, this fraction is shown in Fig. 4 (as a function of phenological phase) and 5 (in the form of a map). In some cases parts of the series could be tested, but not the entire series.

Figure 4 shows for which parameters it was easier to find reference series (left-hand side of the plot) and for which it was more difficult (right-hand side). In general, late phases (fruit maturity, leaf colouring, leaf drop)

have lower spatial correlation and are observed at fewer stations than spring phases, therefore it is harder to find suitable reference series for them.

Concerning species, we find that the cherry tree (*Prunus avium*), the pear tree (*Pyrus communis*), the apple tree (*Malus domestica*) and the dandelion (*Taraxacum officinale*) have enough reference series in almost all stations (not shown).

Figure 5 shows the geographical picture of the breakpoint detection feasibility. Here we clearly see that the mountainous regions (Jura to the north-west and Alps to the south) are those where finding reference series is more difficult. Even on the Swiss Plateau, though, some stations have one quarter of the parameters with insufficient reference series. Taken together, the breakpoint detection was applied to 73.9% (2,925) of the phenological series with at least 20 observations (thereof, 2,082 entire series where subjected to the break detection, for 843 series only segments could be used). For the remaining 26.1% with at least 20 observations (1,035) it was not possible to find enough suitable reference series.

3.3. Number of breakpoints and comparison with metadata

In this section we analyse breakpoints, their attribution to metadata, and their size. Before analysing the statistics of the accepted breakpoints, an example of results for breakpoint detection for one phase of one species at one site is shown in Fig. 6 (see also Fig. S1 and S2 in the Supplementary Material for further examples). The example is for the full flowering of the horse chestnut (*Aesculus hippocastanum*) in Altdorf. In this example, correlations with reference series are high (up to 0.8) and all reference series comprise the entire 35 yrs covered by the candidate. One significant breakpoint is detected in 1995. This breakpoint was detected by two of the three tests (SNHT and Pettitt) by three reference series each. Hence, the breakpoint is barely significant by our definition. However, it can be related to a change of observer and it is followed by a gap of 3 years. The second panel in Figure 6 shows that until 1995 the flowering in Altdorf was usually among the latest, while after 1995 it is often the earliest (except for the last few years). A similar breakpoint (not shown) corresponding to the same change of observer was detected at the same station for the full flowering of the European elder (*Sambucus nigra*) and of the field daisy (*Leucanthemum vulgare*), the latter with the highest possible significance (all reference series in all tests saw the breakpoint); this adds confidence that a change of observer did cause inhomogeneities in the Altdorf series in 1995.

In total 156 breakpoints were detected in 2,925 analysed series. Multiple breakpoints were detected in only 3 series. Figure 7 shows the occurrence of breakpoints for each year in all tested series, as well as the occurrence of changes of observer. Aside from the 1950s (where the low number of stations inflates the frequency of changes), the number of changes is particularly high at the end of the 1980s, in the mid-1990s and in the late 2000s. As one would expect, the main peaks in the occurrence of breakpoints are close to those in the changes of observer. The 1950s are again a special case: here no breakpoints were detected, because the quantity of data is too small and often not enough suitable reference series can be found. Similarly, in the 2000s the breakpoint detection works less well, because the sample after the breakpoint is too small. The peaks in 1987 and 1995 are mostly the result of noise, specifically of simultaneous breakpoints in a few stations caused by new observers (4 breakpoints in the same station are detected in 1987, accounting for nearly 30% of the breakpoints detected in the whole dataset in that year). It is also important to remark that breakpoints near the beginning and the end of a

series are more difficult to detect, so those in the 1980s and 1990s are more often detected also because those years are often in the middle of long series (1986 and 1993 are the two most common middle years).

We estimated the size of each inhomogeneity from the same five reference series used in the breakpoint detection. Figure 8 shows the distribution of the absolute (left panel) and standardized (right panel) sizes of all breakpoints. The distribution is bimodal because breakpoints with a size close to zero are too small to be detected. Moreover, the distribution is not symmetric: there are significantly more negative changes (i.e., anticipation of the phenological phase after the breakpoint) than positive (61%vs. 39%). This asymmetry is found across all phases, although the leaf unfolding is slightly less affected (56% vs. 44%). The reason of the asymmetry is unknown, but our results suggest that it is related to the new observers (the fraction of negative changes reaches 66% for breakpoints confirmed by metadata).

3.4. Performance of the breakpoint detection

All breakpoints detected by the algorithm were confirmed by all the experts, meaning that the false detection rate of the algorithm is virtually zero. In very few cases (3%), however, at least one expert judged the breakpoints to be misplaced by at least two years, or to represent a trend rather than a step function. In 16% of the inhomogeneous series a possible second breakpoint was not detected by the algorithm; moreover in 10% of the randomly selected homogeneous series at least one expert observed at least one breakpoint

From these numbers we extrapolate that 277 additional series, where no breakpoint was detected by the algorithm, contain breakpoints that could be detected visually by an expert. This means that the algorithm is capable of flagging about one third of the series that would be considered inhomogeneous by at least one expert after visual inspection. However, not all of the visually detected breakpoints are real breakpoints. The hit rate could be improved by changing the parameters of the algorithm, but this would come at the cost of more false detections.

3.5. Long phenological series: an example

To illustrate the impact of inhomogeneities on trends, in Fig. 9 we show 10 long series for the full flowering of the dandelion (*Taraxacum officinale*) with similar long-term average, of which one is inhomogeneous. This is an example where a change of observer in the 1980s causes an overestimation of the actual negative trend at one station.

In this subset of stations the flowering of the dandelion changed on average by 13 days over the analysed period. After excluding the inhomogeneous series, the average change is reduced to 11 days.

4. Discussion

Phenological series are less well correlated with each other than temperature series. Particularly in the complex landscape of Switzerland, correlations are relatively low (Güsewell et al. 2018). A minimum correlation of 0.6 was chosen, which for temperature homogenization (based on monthly averages) would be fulfilled by almost all possible station pairs in Switzerland (Gubler et al. 2017). The correlations of phenological series are more similar to those of precipitation, which (for monthly totals in Switzerland) often fall below 0.6 even, sometimes, over short distances. However, in contrast to precipitation, which exhibits correlations close to zero between the

northern and southern slopes of the Alps (Gubler et al. 2017), we did not find a strong influence of the watershed in reducing the correlation.

The period 1986-1989 shows 3 to 4 times more breakpoints than surrounding 4-year periods. This is a period characterized by a rapid temperature increase in Switzerland and similarly rapid changes in phenological variables (Schleip et al. 2008; Reid et al. 2016), but unfortunately it also coincides with a particularly large number of new observers (about 20% of stations affected). If there was an increase of false detections caused by the rapid climate change, we would expect the fraction of breakpoints related to changes of observer to decrease. However, for 1986-1989 this fraction is 57%, even larger than the overall average of 54%.

The estimation of the size of the inhomogeneities is not always reliable. In the example shown in Fig. S1, one breakpoint (1998) barely reaches the significance threshold (three reference series in two tests show a break). A second possible breakpoint in correspondence of a second change of observer in 2001 is only detected by one test and is therefore not significant. Judging from the bottom plot in Fig. S1, the size of the two breakpoints is similar and it is of about one standard deviation; however, since the breakpoint in 2001 was not significant, the whole period 1999-2015 is used to calculate the size of the first breakpoint and this results in an estimated size of only 0.4 standard deviations (i.e., 5 days). There would possibly be a third breakpoint around 1965, but the first 14 years of the series were ignored by the detection algorithm (yellow shading) since not enough reference series were available in that period.

In general, the closer to each other two breakpoints are, the more difficult it becomes to detect them, because of the reduced sample. Moreover, the nature of the tests implies that single breakpoints are much more likely to be detected than multiple breakpoints, because the tests can only detect one breakpoint at a time.

We summarize that breakpoint detection in combination with a thorough analysis of metadata such as observer changes may contribute to a complementary, more robust estimation of breakpoints and shifts in phenological records. Additionally further metadata will be helpful, like e.g. changes in observed plants. An important aspect is the careful quality control of the data, since it has been shown for temperature and precipitation that breakpoint detection and homogenization produces much better results on carefully quality-controlled data (Hunziker et al. 2018).

5. Conclusions

In this paper we have applied breakpoint detection as it is typically used for meteorological variables to phenological data from the Swiss Phenology Network (SPN). We use a combination of three breakpoint tests, each employing 5 reference series that are sufficiently correlated with the candidate. Only a fraction of the series has enough references − spring phases more often than fall phases. The detected breakpoints were compared with metadata, and 54% of them could be attributed to observer changes. We found that especially those breakpoints caused by new observers are more frequently linked to an anticipation of the phenological phases, which is the same signal caused by a warming climate. More detailed metadata, in particular if an observer change corresponds to a change in the observed plants, would help understanding the causes of the asymmetry of the distribution of the breakpoint's sizes and allow the implementation of measures to reduce the systematic impact on trends. However, additional studies are required to assess whether this asymmetry is a special feature of the SPN or whether it can be found in other networks as well. In general, our study demonstrates that the breakpoint

detection methods routinely used in climatology are applicable for phenological series. With this automated method, the fraction of detected breakpoints is comparable with the automated detected fraction for precipitation series. Due to high variability and comparably low correlations between phenological series, mainly large breakpoints are detected. The graphical and statistical outputs of the breakpoint detection can be used for further assessment of the most valuable data series by a manual control.

In the future, citizen science data might be used more frequently for gaining phenological data (Lehmann et al. 2018), which makes quality control and the assessment of long-term stability even more important.

Acknowledgements

This work was funded by MeteoSwiss in the framework of GCOS Switzerland**.**

References

Alexandersson H, Moberg A (1997) Homogenization of Swedish temperature data. Part I: Homogeneity test for linear trends. Int J Climatol 17: 25-34. https://doi.org/10.1002/(SICI)1097-0088(199701)17:1<25::AID-JOC103>3.0.CO;2-J

Anderson DM, Mauk EM, Wahl ER, Morrill C, Wagner AJ, Easterling D, RutishauserT (2013) Global warming in an independent record of the past 130 years. Geophys Res Lett 40: 189–193. https://doi.org/10.1029/2012GL054271

Auchmann R, Brugnara Y, Rutishauser T, Brönnimann S, Gehrig R, Pietragalla B, Begert M, Sigg C, Knechtl V, Calpini B, Konzelmann T (2018) Quality analysis and classification of data series from the Swiss Phenology Network. Technical Report MeteoSwiss 269, MeteoSwiss, Zurich

Begert M, Schlegel T, Kirchhofer W (2005) Homogeneous temperature and precipitation series of Switzerland from 1864 to 2000. Int J Climatol 25: 65-80. https://doi.org/10.1002/joc.1118Brügger R, Vassella A (2003) Pflanzen im Wandel der Jahreszeiten - Anleitung für phänologische Beobachtungen. Geographica Bernensia, Bern

Cramer W, Yohe GW, Auffhammer M, Huggel C, Molau U, da Silva Dias, MAF, Solow A, Stone DA, Tibig L (2014) Detection and attribution of observed impacts. In: Field CB, Barros VR, Dokken DJ et al (eds) Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA

Čufar K, De Luis M, Saz MA, Črepinšek Z, Kajfež-Bogataj L (2012) Temporal shifts in leaf phenology of beech (Fagus sylvatica) depend on elevation. Trees 26: 1091-1100. https://doi.org/10.1007/s00468-012-0686-7

Defila C, Clot B (2001) Phytophenological trends in Switzerland. Int J Biometeorol 45: 203–207. https://doi.org/10.1007/s004840100101

Defila C (2008) Plant phenological observations in Switzerland. In: Nekovář J, Koch E, Kubin E, Nejedlik P, Sparks T, Wielgolaski FE (eds) The history and current status of plant phenology in Europe.COST office, Bruxelles

Defila C, Clot B, Jeanneret F, Stöckli R (2016) Phenology in Switzerland since 1808. In: Willemses S, Fuger M (eds) From weather observations to atmospheric and climate sciences in Switzerland. vdf Hochschulverlag, Zurich

Dose V, Menzel A (2004) Bayesian analysis of climate change impacts in phenology. Global Change Biol 10: 259-272. https://doi.org/10.1111/j.1529-8817.2003.00731.x

Fu YH, Zhao H, Piao S, Peaucelle M, Peng S, Zhou G, Ciais P, Huang M, Menzel A, Peñuelas J, Song Y (2015) Declining global warming effects on the phenology of spring leaf unfolding. Nature 526: 104. https://doi.org/10.1038/nature15402

Ge Q, Wang H, Zheng J, Rutishauser T, Dai J (2014) A 170 year spring phenology index of plants in eastern China. J Geophys Res 119: 301-311. https://doi.org/10.1002/2013JG002565

Gubler S, Hunziker S, Begert M, Croci-Maspoli M, Konzelmann T, Brönnimann S, Schwierz C, Oria C, Rosas G (2017) The influence of station density on climate data homogenization using HOMER. Int J Climatol 37: 4670–4683. https://doi.org/10.1002/joc.5114

Güsewell S, Pietragalla B, Gehrig R, Furrer R (2018) Representativeness of stations and reliability of data in the Swiss Phenology Network. Technical Report MeteoSwiss 267, MeteoSwiss, Zurich

Hense A, Müller M (2007) Geostatistische Modellierung und Qualitätskontrolle von phänologischen Beobachtungen. Promet 33: 2–6

Hunziker S, Brönnimann S, Calle J, Moreno I, Andrade M, Ticona L Lavado W, Huerta A (2018) Effects of undetected data quality issues on climatological analyses. Clim Past 14: 1-20. https://doi.org/10.5194/cp-14-1-2018

Jamali S, Jönsson P, Eklundh L, Ardö J, Seaquist J (2015) Detecting changes in vegetation trends using time series segmentation. Remote Sens Environ 156: 182-195. https://doi.org/10.1016/j.rse.2014.09.010

Keatley MR, Hudson IL (2012) Detecting change in an Australian flowering record: Comparisons of linear regression and CUSUM change point analysis. Austral Ecol 37:825-835. https://doi.org/10.1111/j.1442-9993.2011.02344.x

Kuglitsch FG, Auchmann R, Bleisch R, Brönnimann S, Martius O, Stewart M (2012) Break detection of annual Swiss temperature series. J Geophys Res 117: D13105. https://doi.org/10.1029/2012JD017729

Lehmann D, Wyss E, Rutishauser T, Brönnimann S (2018) Citizen Science: Pflanzenphänologische Daten erfüllen wissenschaftliche Kriterien. Geographica Bernensia G93: 1-4. https://doi.org/10.4480/GB2017.G93Menzel A, Sparks TH, Estrella N et al (2006) European phenological response to climate change matches the warming pattern. Global Change Biol 12: 1969–1976. https://doi.org/10.1111/j.1365-2486.2006.01193.x

MeteoSwiss (2018) National Climate Observing System (GCOS Switzerland). Update 2018, MeteoSwiss, Zurich. https://www.meteoswiss.admin.ch/content/dam/meteoswiss/en/Forschung-und-

Zusammenarbeit/Internationale-Zusammenarbeit/GCOS/doc/NationalClimateObservingSystem_GCOSSwitzerland-Small.pdf. Accessed 11 August 2019

Pettitt AN (1979) A non-parametric approach to the change-point detection. Appl Statist 28: 126-135

Reid PC, Hari RE, Beaugrand G et al (2016) Global impacts of the 1980s regime shift. Glob Chang Biol 22: 682–703. https://doi.org/10.1111/gcb.13106.51

Rutishauser T, Luterbacher J, Defila C, Frank D, Wanner H (2008) Swiss Spring Plant Phenology 2007: Extremes, a multi-century perspective and changes in temperature sensitivity. Geophys Res Lett 35: L05703. https://doi.org/10.1029/2007GL032545

Schleip C, Rutishauser T, Luterbacher J, Menzel A (2008) Time series modeling and central European temperature impact assessment of phenological records over the last 250 years. J Geophys Res 113: G04026. https://doi.org/10.1029/2007JG000646

Seiz G, Foppa N (2007) Nationales Klima-Beobachtungssystem (GCOS Schweiz). MeteoSwiss, Zurich. https://www.meteoschweiz.admin.ch/content/dam/meteoswiss/de/Ungebundene-Seiten/Publikationen/Berichte-GCOS-Schweiz/doc/2007seizfoppa.pdf. Accessed 11 August 2019

Studer S, Appenzeller C, Defila C (2005) Inter-annual variability and decadal trends in Alpine spring phenology: A multivariate analysis approach. Clim Change 73: 395–414. https://doi.org/10.1007/s10584-005-6886-z

Venema VKC, Mestre O, Aguilar E et al (2012) Benchmarking homogenization algorithms for monthly data. Clim Past 8: 89-115. https://doi.org/10.5194/cp-8-89-2012

Verbesselt J, Hyndman R, Zeileis A, Culvenor D (2010) Phenological change detection while accounting for abrupt and gradual trends in satellite image time series. Remote Sens Environ 114: 2970-2980. https://doi.org/10.1016/j.rse.2010.08.003

Wang XL, Wen QH, Wu Y (2007) Penalized maximal t test for detecting undocumented mean change in climate data series. J Appl Meteorol Clim 46: 916-931. https://doi.org/10.1175/JAM2504.1

Wang XL (2008) Penalized maximal F test for detecting undocumented mean shift without trend change, J Atmos Oceanic Technol 25: 368–384. https://doi.org/10.1175/2007JTECHA982.1

Willett KM, Williams C, Jolliffe IT et al (2014) A framework for benchmarking of homogenisation algorithm performance on the global scale. Geosci Instrum Method Data Syst 3: 187-200. https://doi.org/10.5194/gi-3-187-2014

WMO (2016) The global observing system for climate: implementation needs. GCOS-200. World Meteorological Organization, Geneva

Table 1: Settings and outcome of different breakpoint detection experiments. Biological constraint: reference must be same species and phase of the candidate (exceptions are start of flowering/ full flowering and the leaf or needle colouring/leaf or needle drop). Statistical tests: number of significant tests out of 3 tests required to accept a breakpoint. Reference series: number of reference series. Max onset diff: the phenological mean onset cannot differ more than a certain period between candidate and reference series. Min correlation: minimum Pearson's correlation of reference series. Tolerance: tolerance in years for detected breakpoints. Quality controlled: data underwent quality control before breakpoint detection. If more than the maximum number of series fulfil all requirements those with most overlapping observations were considered. The chosen configuration is highlighted in bold.

	ALL1	ALL2	8REF	**5REF**	5REF_NOQC
biological constraint	No	No	Yes	**Yes**	Yes
statistical tests	2/3	2/3	2/3	**2/3**	2/3
reference series	8	8	8	**5**	5
min overlap	No	No	90%	**90%**	90%
min length (yrs)	10	10	20	**20**	20
max elevation diff (m)	1000	1000	750	**750**	750
max onset diff (days)	No	No	30	**30**	30
min correlation	0.6	0.6	0.6	**0.6**	0.6
tolerance (yrs)	1	2	1	**1**	1
quality controlled	Yes	Yes	Yes	**Yes**	No
series tested	7393	7393	2566	**2925**	2951
breakpoints	485	644	330	**156**	141
- confirmed by metadata	209 (43%)	257 (40%)	140 (42%)	**85 (54%)**	76 (54%)

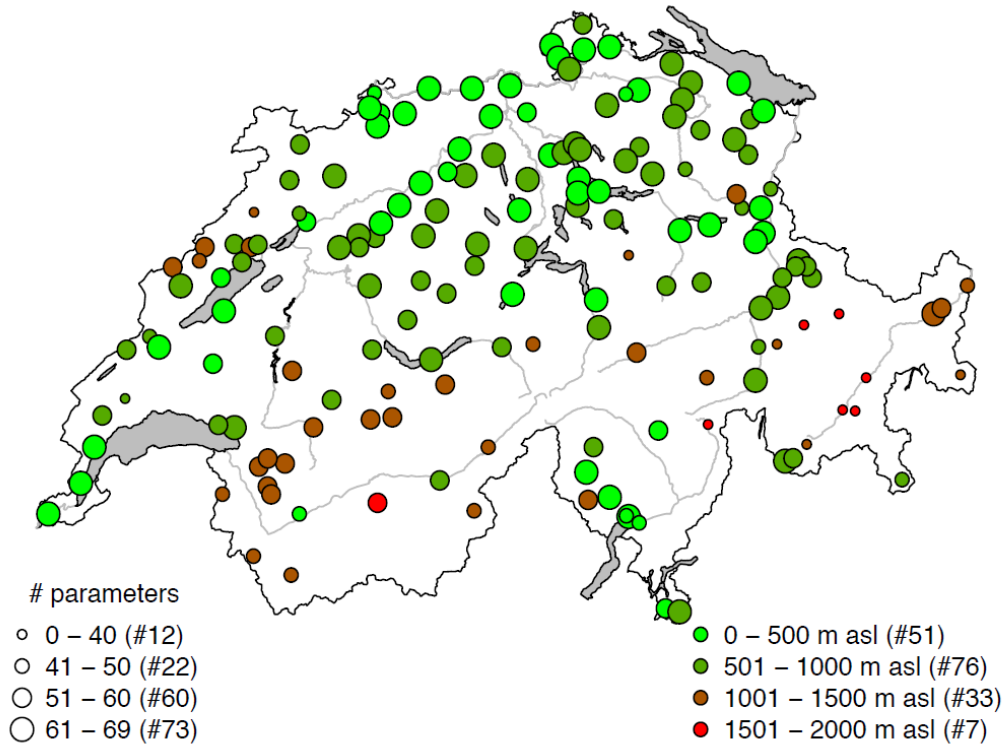

Figure 1: Location of phenology stations by elevation(colors) and number of parameters per station (point size). The left legend shows in brackets the number of stations for each number-of-parameters class.

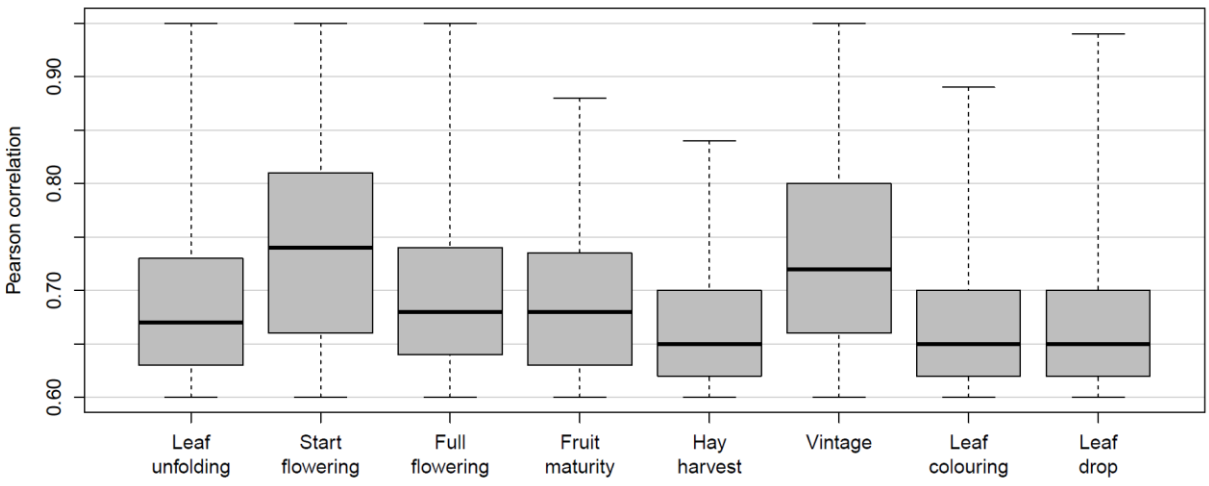

Figure 2: Boxplots of the Pearson correlation of the reference series for each phase. The whiskers encompass the full range of values. Reference series have a minimal correlation of 0.6 according to our criterion. Note that the needle emergence/colouring/drop are included into leaf unfolding/colouring/drop.

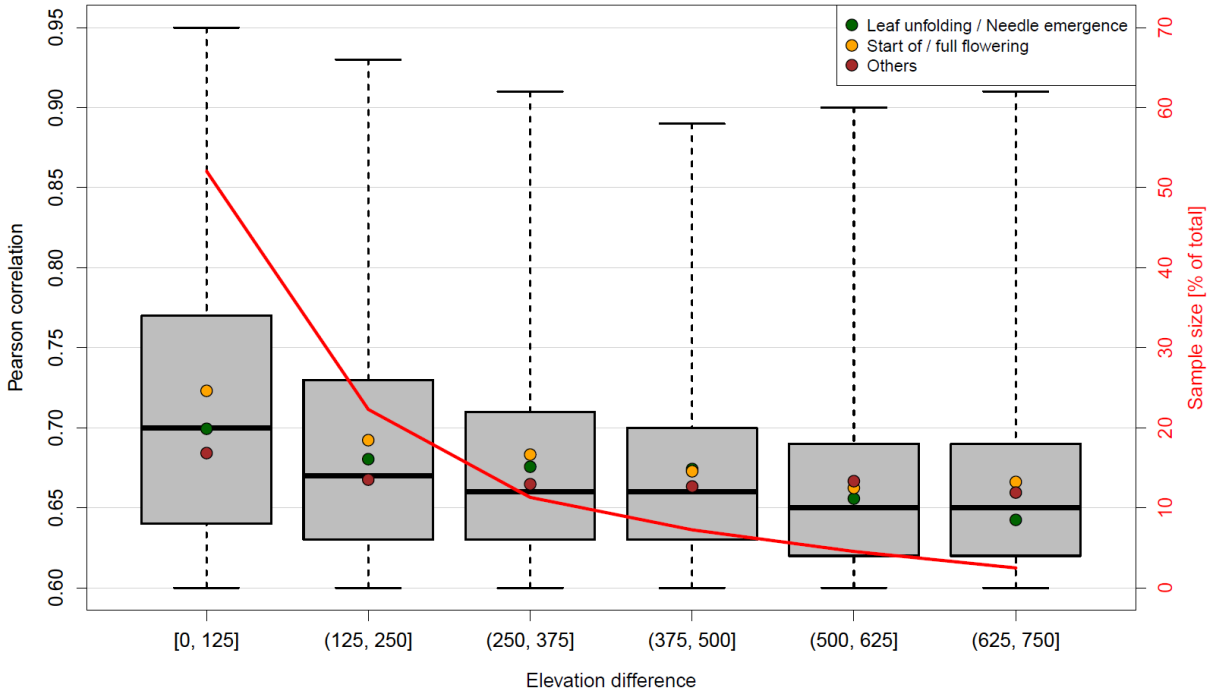

Figure 3: Boxplots of the correlations between candidate and reference series separated by elevation differences. The points show the averages of different phases, the red lines show how many reference series contributed to each boxplot.

5

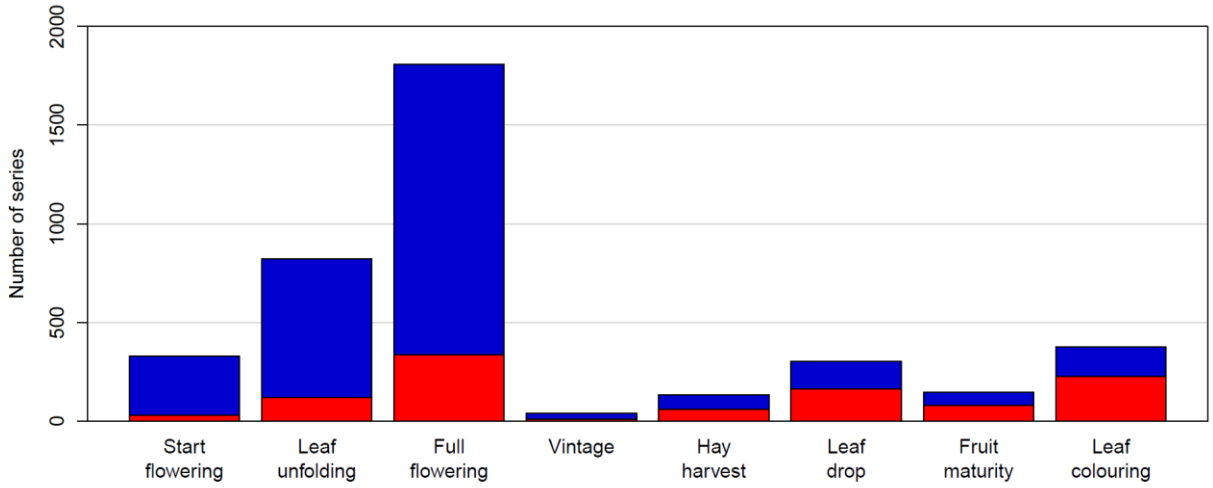

Figure 4: Breakpoint detection feasibility for each phase. The red fraction of the bars is the fraction of series that did not have enough reference series for any period of at least 20 years. Note that the needle emergence/colouring/drop are included into leaf unfolding/colouring/drop.

10

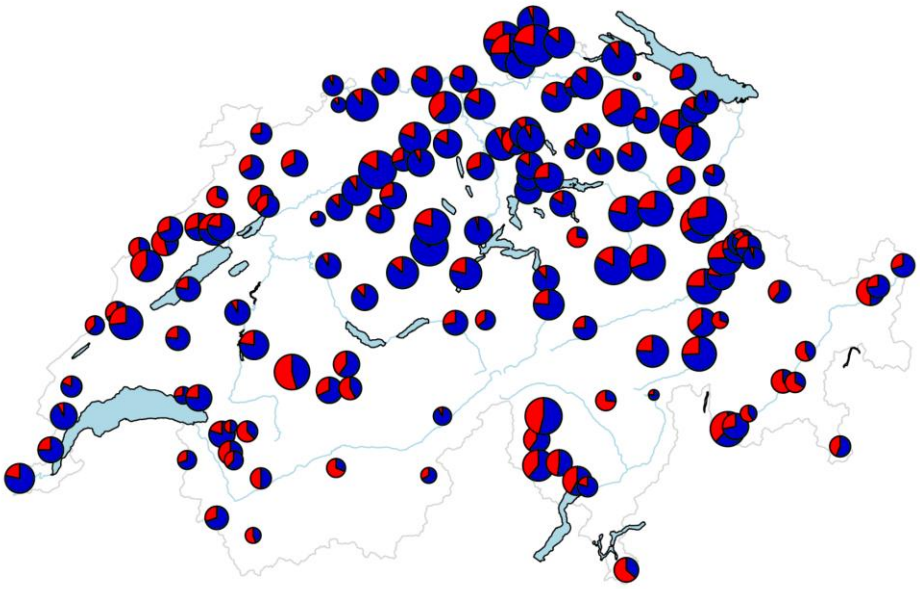

Figure 5: Breakpoint detection feasibility for each station. The red fraction of the pies is the fraction of series that did not have enough reference series in any period of at least 20 years. The area of the pies is proportional to the number of parameters.

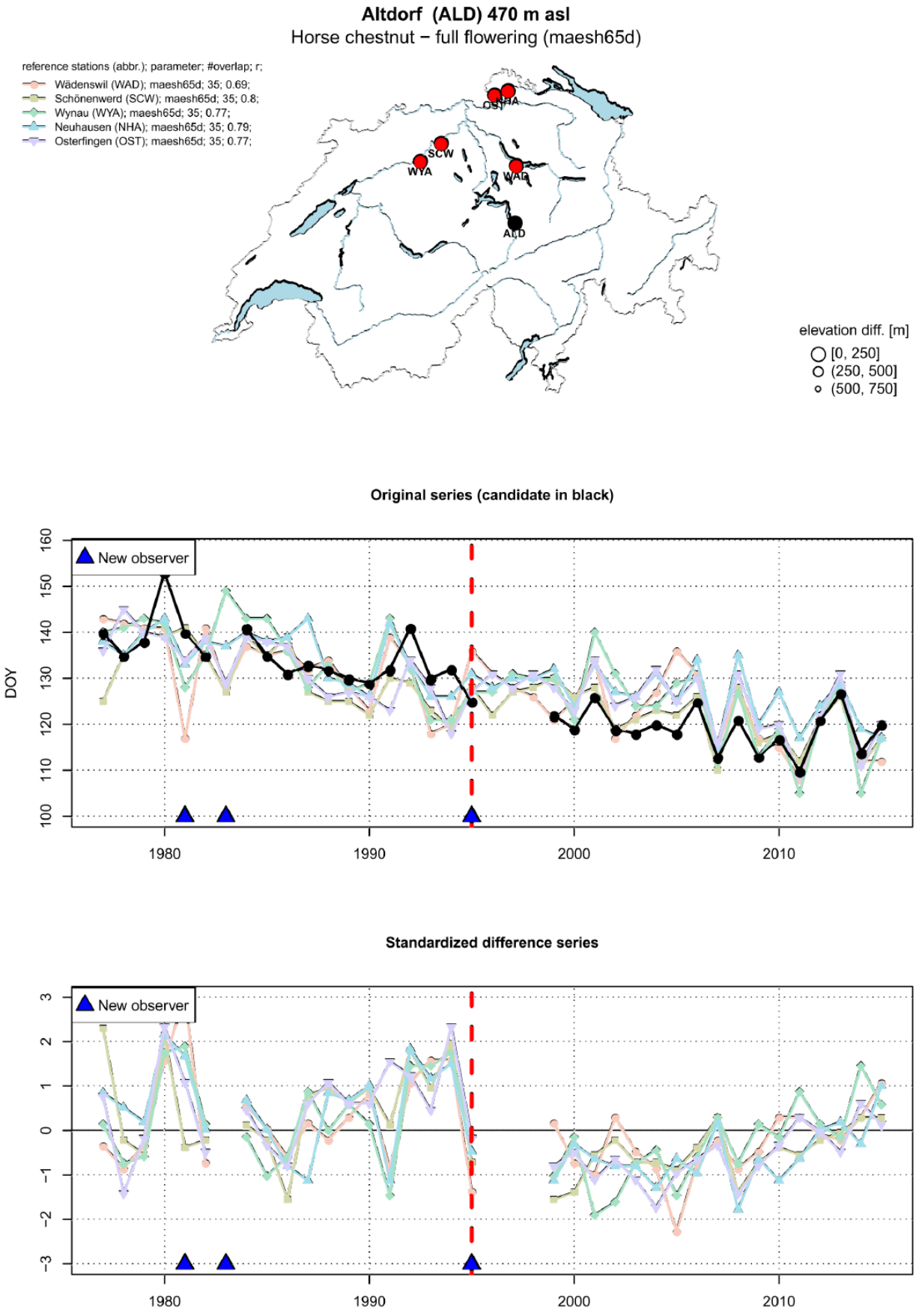

Figure 6: Breakpoint detection summary plot for the full flowering of the horse chestnut in Altdorf. The map shows the position of the candidate (black point) and reference series (red points), the vertical dashed line in the time series indicates the position of the breakpoint and the triangles the changes of observer.

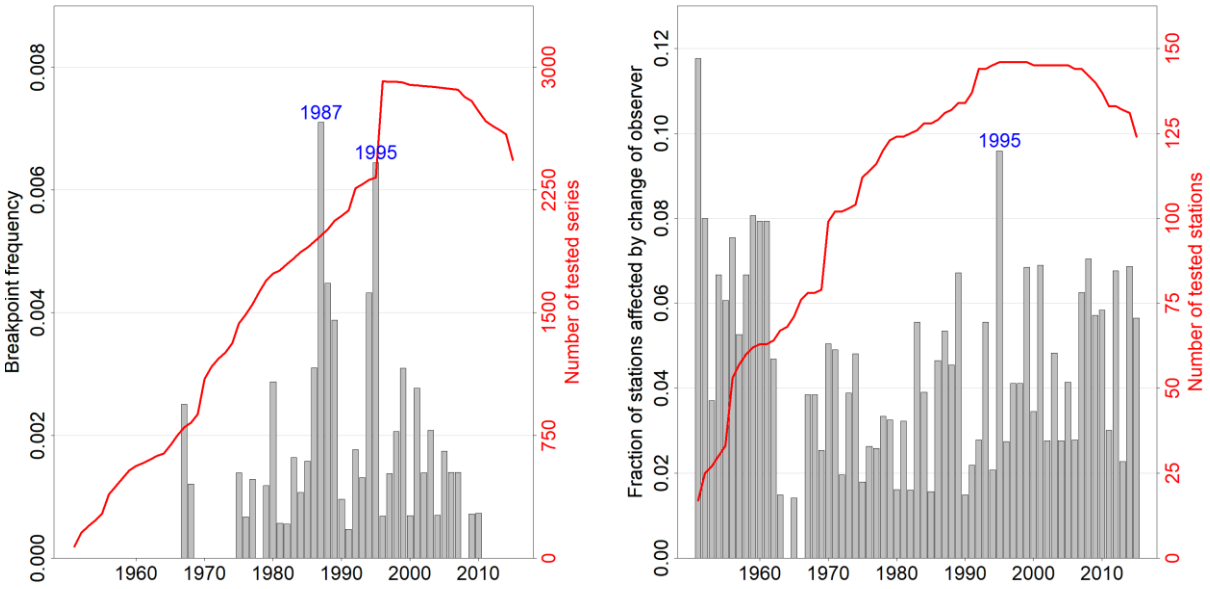

Figure 7: (left) Number of significant breakpoints detected (relative to the number of tested series). (right) Number of changes of observer (relative to the number of tested stations). The red lines depict the number of tested series / stations.

5

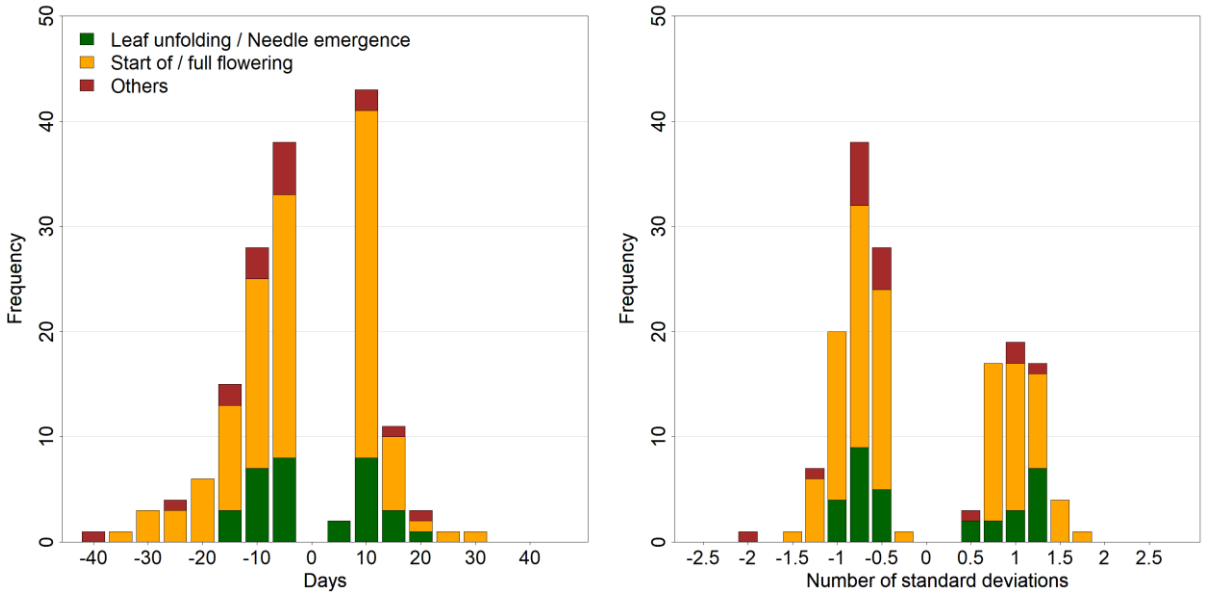

Figure 8: Histograms of the absolute (left) and standardized (right) size of the detected inhomogeneities.

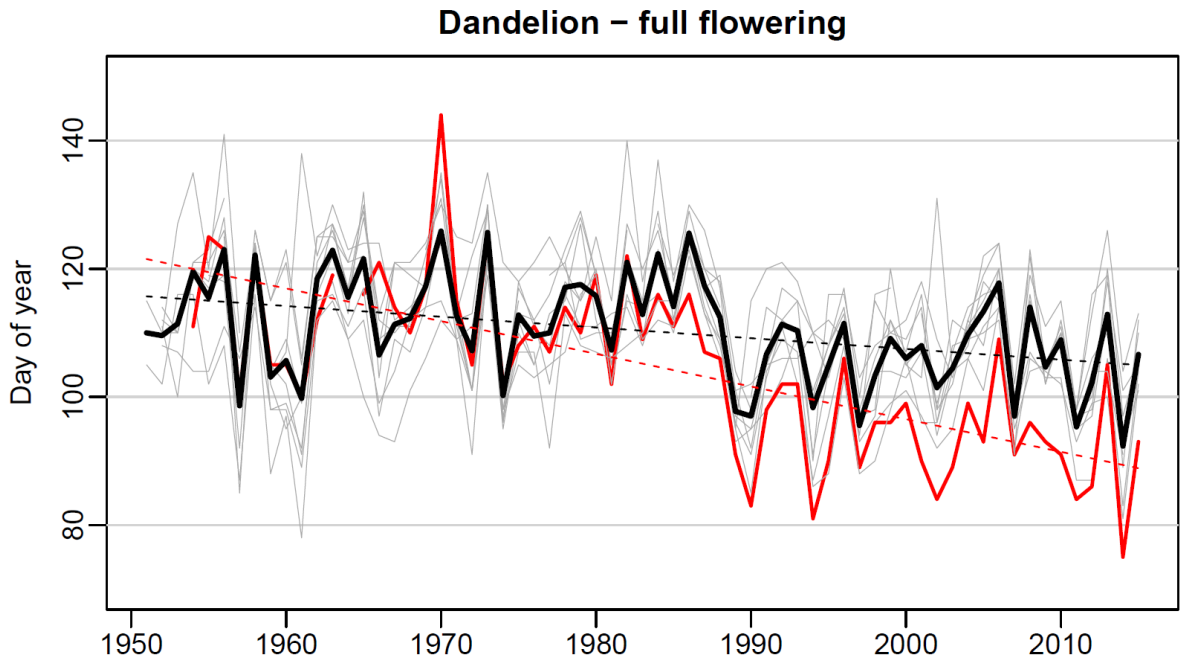

Figure 9: Selected time series of the full flowering of the dandelion (Taraxacum officinale). The red line represents a series with a breakpoint in 1986, the black line represents the average of all the other series. Dashed lines are the respective linear trends.

5